Jonathan Skaggs

CS 478 Clustering

HAC Sponge (single link)

Number of Clusters : 4

Cluster: [0, 49, 61, 52, 12, 60, 63, 64, 27, 40, 73, 53, 72, 70, 71, 30, 32, 21, 23, 1, 7, 33, 50, 69, 2, 4, 5, 8, 9, 42, 38, 54, 55, 56, 34, 37, 44, 45, 68, 10, 35, 59, 11, 43, 51, 66, 67, 57, 58, 62, 65, 3, 13, 47, 48, 14, 20, 24, 29, 74, 75, 16, 25, 31, 26, 28, 19, 15, 17, 22, 36, 39, 41]

Centroid Values: ['2_CAPAS', 'SIN_CAPA_INTERNA_DEL_CORTEX', 'NO', 1.521, 'INTERMEDIARIOS', 'NORMAL', 2.2603, 0.932, 'OTROS']

Size: 73

SSE: 604.931506849

Cluster: [6]

Centroid Values: ['1_CAPA', 'SIN_CAPA_INTERNA_DEL_CORTEX', 'SI', 4.0, 'SIN_TILOSTILOS_ADICIONALES', 'SIN_ESPICULA_PRINCIPAL_ESTILO', 1.0, 3.0, 'OTROS']

Size: 1

SSE: 0.0

Cluster: [18]

Centroid Values: ['3_CAPAS', 'TANGENCIAL', 'SI', 3.0, 'ECTOSOMICOS_DISPERSOS', 'FUSIFORME', 3.0, 4.0, ?]

Size: 1

SSE: 1.0

Cluster: [46]

Centroid Values: ['3_CAPAS', 'TANGENCIAL', 'SI', 2.0, 'SIN_TILOSTILOS_ADICIONALES', 'SIN_ESPICULA_PRINCIPAL_ESTILO', 1.0, 0.0, ?]

Size: 1

SSE: 1.0

Total SSE: 606.932

HAC Sponge (complete link)

Number of Clusters : 4

Cluster: [0, 57, 1, 5, 2, 3, 4, 7, 9, 12, 10, 13, 11, 6, 8]

Centroid Values: [6.0, 2.429, 4.323, 4.727, 4.0, "'none'", 38.308, "'empl_contr'", 12.0, 5.800, "'yes'", 11.929, "'generous'", "'yes'", "'half'", "'yes'", "'half'", "'good'"]

Size: 15

SSE: 964.551

Cluster: [14, 15, 16, 19, 22, 17, 18, 20, 21, 23, 24, 25, 26, 28, 27, 29]

Centroid Values: [21.5, 1.875, 3.481, 3.673, 3.300, "'none'", 36.600, "'none'", 7.667, 4.5, "'yes'", 10.933, "'below_average'", "'yes'", "'half'", "'yes'", "'full'", "'bad'"]

Size: 16

SSE: 826.896

Cluster: [30, 33, 31, 32, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45]

Centroid Values: [37.5, 2.063, 3.0500, 3.192, 2.900, "'none'", 39.333, "'none'", 2.66, 2.625, "'no'", 10.625, "'below_average'", "'no'", "'none'", "'yes'", "'none'", "'bad'"]

Size: 16

SSE: 538.852

Cluster: [46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56]

Centroid Values: [51.0, 2.545, 4.755, 4.436, 4.5, "'none'", 38.222, "'empl_contr'", 12.0, 7.200, "'yes'", 11.111, "'average'", "'yes'", "'full'", "'yes'", "'full'", "'bad'"]

Size: 11

SSE: 333.824

Total SSE: 2664.123

K-Means Sponge

Iteration 0

Centroid: [0.0, 1.0, 5.0, ?, ?, ?, 40.0, ?, ?, 2.0, ?, 11.0, 1.0, ?, ?, 0.0, ?, 1.0]

Centroid: [1.0, 2.0, 4.5, 5.8, ?, ?, 35.0, 1.0, ?, ?, 0.0, 11.0, 0.0, ?, 2.0, ?, 2.0, 1.0]

Centroid: [2.0, ?, ?, ?, ?, ?, 38.0, 2.0, ?, 5.0, ?, 11.0, 2.0, 0.0, 1.0, 0.0, 1.0, 1.0,]

Centroid: [3.0, 3.0, 3.7, 4.0, 5.0, 2.0, ?, ?, ?, ? , 0.0, ?, ?, ?, ?, 0.0, ?, 1.0]

SSE: 52027.89

Iteration 1

Centroid: [-0.5, 2.5, 5.0, ?, 1.0, 1, 40.5, 0, ?, 2.0, 0, 12.0, 1, 0, 0, 0, 0, 1]

Centroid: [1. 2. 4.5 5.8 ? ? 35. 1. ? ? 0. 11. 0. ? 2. ? 2. 1.]

Centroid: [2. ? ? ? ? ? 38. 2. ? 5. ? 11. 2. 0. 1. 0. 1. 1.]

Centroid: [29.5, 2.1851, 3.769, 3.931, 3.913, 0, 38.063, 0, 7.444, 4.966, 1, 11.1, 0, 0, 1, 0, 2, 0]

SSE: 9770.933

Iteration 2

Centroid: [-0.5, 2.5, 5.0, ?, 1.0, 1, 40.5, 0, ?, 2.0, 0, 12.0, 1, 0, 0, 0, 0, 1]

Centroid: [4.5, 2.0, 3.75, 6.4000000000000004, ?, 0, 36.5, 1, 12.0, 25.0, ?, 11.0, ?, 0, 1, 0, 2, 0]

Centroid: [8.538, 2.333, 4.267, 4.320, 4.5, 0, 37.909, 2, ?, 4.555, 0, 11.917, 2, 0, 2, 0, 1, 0]

Centroid: [36.0, 2.146, 3.641, 3.726, 3.522, 0, 38.108, 0, 6.875, 4.150, 1, 10.842, 0, 0, 1, 0, 2, 0]

SSE: 6070.62831325

Iteration 3

Centroid: [0.333, 2.5, 5.0, ?, 1.0, 1, 39.666, 2, ?, 3.5, 0, 11.666, 1, 0, 1, 0, 1, 1]

Centroid: [4.0, 2.5, 3.925, 5.325, 5.0, 2, 37.667, 1, 12.0, 25.0, 0, 11.333, 0, 0, 1, 0, 1, 0]

Centroid: [13.824, 2.0, 3.776, 4.258, 4.320, 0, 37.0, 2, 7.5, 4.167, 1, 11.875, 2, 0, 2, 0, 2, 0]

Centroid: [39.5, 2.235, 3.768, 3.677, 3.388, 0, 38.567, 0, 6.667, 4.313, 1, 10.688, 1, 0, 1, 0, 2, 0]

SSE: 4648.13655328

Iteration 4

Centroid: [2.600, 2.5, 4.633, 3.650, 1.650, 1, 38.800, 2, ?, 4.0, 0, 11.800, 1, 0, 1, 0, 1, 1]

Centroid: [4.400, 2.600, 3.940, 5.260, 5.0, 2, 37.667, 1, 12.0, 25.0, 0, 11.5, 0, 0, 1, 0, 1, 0]
Centroid: [17.5, 1.889, 3.794, 4.108, 4.075, 0, 37.353, 0, 7.667, 4.0, 1, 11.353, 2, 0, 2, 0, 2, 0]
Centroid: [41.5, 2.267, 3.703, 3.696, 3.586, 0, 38.481, 0, 6.400, 4.429, 1, 10.821, 1, 0, 1, 0, 2, 0]
SSE: 3972.582

Iteration 5
Centroid: [3.667, 2.200, 4.900, 3.650, 1.650, 0, 39.0, 2, ?, 4.0, 0, 11.667, 1, 0, 1, 0, 1, 1]
Centroid: [4.400, 2.600, 3.940, 5.260, 5.0, 2, 37.667, 1, 12.0, 25.0, 0, 11.5, 0, 0, 1, 0, 1, 0]
Centroid: [19.5, 2.0, 3.630, 3.920, 4.075, 0, 36.842, 0, 7.667, 4.083, 1, 11.315, 2, 0, 2, 0, 2, 0]
Centroid: [43.0, 2.259, 3.744, 3.763, 3.586, 0, 38.917, 2, 6.400, 4.385, 1, 10.800, 1, 0, 1, 0, 2, 0]
SSE: 3705.989

Iteration 6
Centroid: [5.375, 2.286, 4.917, 4.425, 2.633, 0, 38.5, 2, ?, 3.857, 0, 12.25, 2, 0, 2, 0, 1, 1]
Centroid: [4.400, 2.600, 3.940, 5.260, 5.0, 2, 37.667, 1, 12.0, 25.0, 0, 11.5, 0, 0, 1, 0, 1, 0]
Centroid: [21.0, 2.0, 3.537, 3.779, 4.175, 0, 37.0, 0, 7.667, 4.200, 1, 11.0, 0, 0, 1, 0, 2, 0]
Centroid: [43.5, 2.231, 3.715, 3.730, 3.350, 0, 38.869, 2, 6.400, 4.385, 1, 10.792, 1, 0, 1, 0, 2, 0]
SSE: 3522.269

Iteration 7
Centroid: [6.800, 2.333, 4.563, 4.283, 3.25, 0, 38.5, 2, ?, 3.667, 0, 12.1, 2, 0, 2, 0, 1, 1]
Centroid: [4.400, 2.600, 3.940, 5.260, 5.0, 2, 37.667, 1, 12.0, 25.0, 0, 11.5, 0, 0, 1, 0, 1, 0]
Centroid: [23.0, 2.0, 3.458, 3.529, 3.525, 0, 37.056, 0, 7.667, 4.555, 1, 10.833, 0, 0, 1, 0, 2, 0]
Centroid: [44.5, 2.208, 3.796, 3.871, 3.520, 0, 38.857, 2, 6.400, 4.333, 1, 10.864, 1, 0, 0, 0, 2, 0]
SSE: 3343.618

Iteration 8
Centroid: [8.090, 2.100, 4.350, 4.386, 4.25, 0, 38.0, 2, ?, 4.300, 1, 11.909, 2, 0, 2, 0, 1, 0]
Centroid: [3.0, 2.800, 3.9250, 5.325, 3.667, 1, 38.5, 1, 12.0, 25.0, 0, 11.75, 0, 0, 1, 0, 1, 1]
Centroid: [24.0, 2.053, 3.511, 3.627, 3.525, 0, 37.278, 0, 7.667, 3.778, 1, 10.778, 0, 0, 1, 0, 2, 0]
Centroid: [45.0, 2.217, 3.78, 3.815, 3.520, 0, 38.800, 2, 6.400, 4.455, 1, 10.905, 1, 0, 1, 0, 2, 0]
SSE: 3221.923

Iteration 9
Centroid: [8.667, 2.091, 4.364, 4.338, 4.25, 0, 37.909, 2, ?, 4.300, 1, 11.833, 2, 0, 2, 0, 1, 0]
Centroid: [3.0, 2.800, 3.925, 5.325, 3.667, 1, 38.5, 1, 12.0, 25.0, 0, 11.75, 0, 0, 1, 0, 1, 1]
Centroid: [25.0, 2.105, 3.379, 3.527, 3.240, 0, 37.444, 0, 6.25, 3.5, 1, 10.722, 0, 0, 1, 0, 2, 0]
Centroid: [45.5, 2.182, 3.868, 3.884, 3.875, 0, 38.737, 2, 7.5, 4.800, 1, 10.950, 1, 0, 2, 0, 2, 0]
SSE: 3144.943

Iteration 10
Centroid: [9.786, 1.923, 4.069, 4.338, 4.25, 0, 37.846, 2, 2.0, 4.0, 1, 11.643, 2, 0, 2, 0, 1, 0]

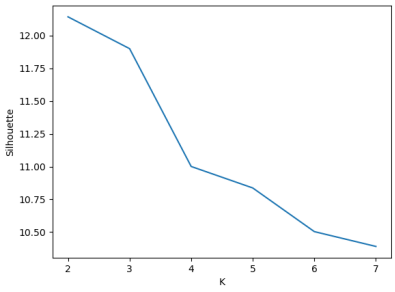
Centroid: [3.0, 2.800, 3.925, 5.325, 3.667, 1, 38.5, 1, 12.0, 25.0, 0, 11.75, 0, 0, 1, 0, 1, 1]
Centroid: [26.5, 2.222, 3.406, 3.431, 3.240, 0, 37.588, 0, 7.667, 3.75, 1, 10.765, 0, 0, 1, 0, 2, 0]
Centroid: [46.0, 2.190, 3.957, 3.989, 3.875, 0, 38.667, 2, 7.5, 4.800, 0, 10.947, 1, 0, 2, 0, 2, 0]
SSE: 3047.0146

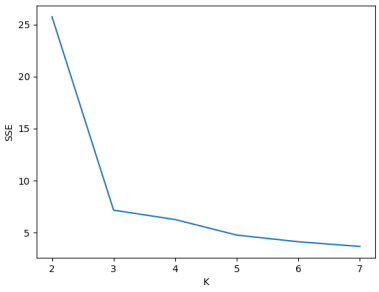
Iteration 11
Centroid: [10.642, 1.769, 3.915, 4.243, 4.0, 0, 37.857, 2, 2.0, 4.0, 1, 11.571, 2, 0, 2, 0, 0, 0]
Centroid: [3.5, 2.833, 3.940, 5.260, 4.0, 2, 38.5, 1, 12.0, 25.0, 0, 11.800, 0, 0, 1, 0, 1, 1]
Centroid: [27.5, 2.222, 3.406, 3.431, 3.240, 0, 37.706, 0, 6.75, 3.333, 1, 10.765, 0, 0, 1, 0, 2, 0]
Centroid: [46.5, 2.25, 4.0550, 3.989, 3.875, 0, 38.588, 2, 8.667, 5.333, 0, 10.944, 1, 0, 2, 0, 2, 0]
SSE: 2966.216

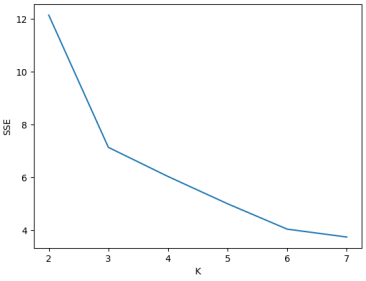
Iteration 12
Centroid: [10.643, 1.769, 3.915, 4.242, 4.0, 0, 37.857, 2, 2.0, 4.0, 1, 11.571, 2, 0, 2, 0, 0, 0]
Centroid: [3.5, 2.833, 3.940, 5.260, 4.0, 2, 38.5, 1, 12.0, 25.0, 0, 11.800, 0, 0, 1, 0, 1, 1]
Centroid: [28.0, 2.158, 3.373, 3.431, 3.240, 0, 37.722, 0, 5.800, 3.300, 1, 10.667, 0, 0, 1, 0, 2, 0]
Centroid: [47.0, 2.316, 4.121, 3.989, 3.875, 0, 38.625, 2, 12.0, 5.625, 0, 11.059, 1, 0, 2, 0, 2, 0]
SSE: 2928.605

Iteration 13
Centroid: [10.643, 1.769, 3.915, 4.242, 4.0, 0, 37.857, 2, 2.0, 4.0, 1, 11.571, 2, 0, 2, 0, 0, 0]
Centroid: [3.5, 2.833, 3.940, 5.260, 4.0, 2, 38.5, 1, 12.0, 25.0, 0, 11.800, 0, 0, 1, 0, 1, 1]
Centroid: [28.0, 2.158, 3.373, 3.431, 3.240, 0, 37.722, 0, 5.800, 3.300, 1, 10.667, 0, 0, 1, 0, 2, 0]
Centroid: [47.0, 2.316, 4.121, 3.989, 3.875, 0, 38.625, 2, 12.0, 5.625, 0, 11.059, 1, 0, 2, 0, 2, 0]
SSE: 2928.605

HAC (single / complete) / K-Means Iris (with normalization, without label)



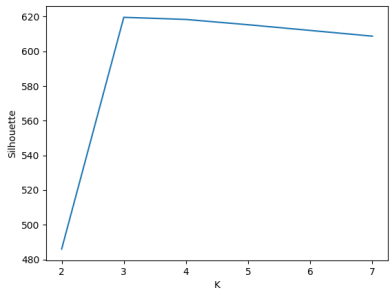


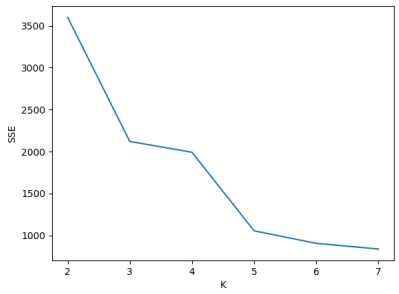


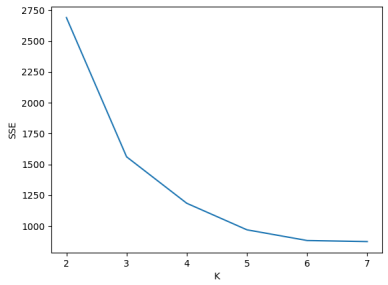
HAC (single / complete) / K-Means Iris (with normalization, with label k=4)

Normalization helps to give each feature an even weighting. Otherwise, the larger numbered variables have a greater impact on the distance metrics. When I ran the algorithms, I was surprised that it got the nearly the same same SSE, centroid values, and clusters results almost every time. The order in which it combine clusters was different but the final clusters were almost always similar. Through the process of the algorithm is seems to less dependent on the ordering of the rows, than I previously thought.

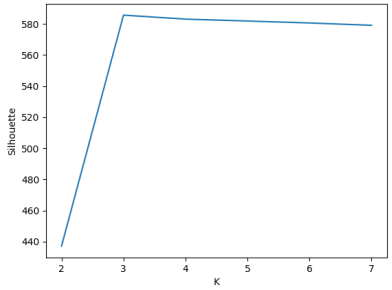
HAC (single / complete) / K-Means Abalone (without normalization)

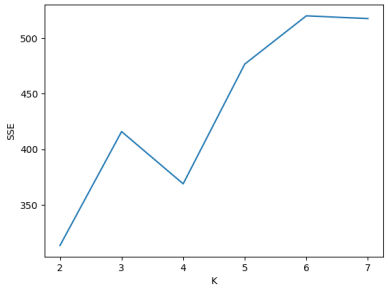


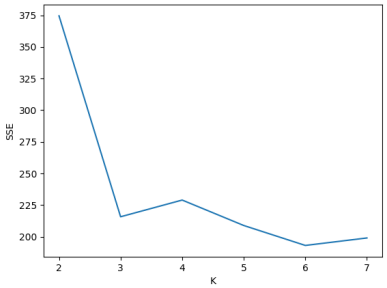




HAC (single / complete) / K-Means Abalone (with normalization)

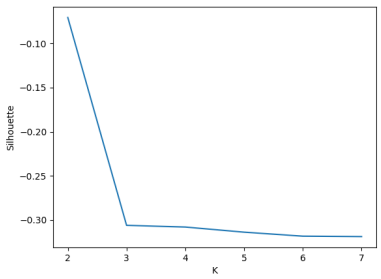


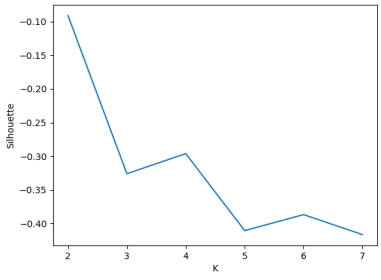


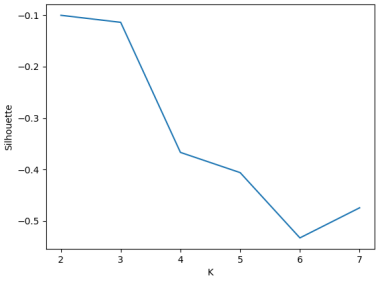


It is better to treat rings as continuous because, although they are always integers, order matters. The number of rings is not a classification therefore letting it use real numbers will help our model predict more accurately.

HAC (single / complete) / K-Means Abalone (silhouette scores, with normalization)







Silhouette scores are not any more indicative of better results than the sum squared error. Sum squared error is more indicative of how close the points are to their centroids. Silhouette scores are more indicative of how much separation there is between clusters. When choosing between using silhouette scores and sum squared error, you need to decide what you care about. Only after this can you decide which to choose. Once you decide which metric to choose then you should look at the metric in order to decide what number you want for k.

Further Experiments

I noticed when we ran the sponge experiments we decided to make all unknown values = 1 what would happen if you set it instead to the largest value for that feature. Would you get better results? My hypothesis is yes. I think that setting the distances to 1 often allows missing values to appear closer than they might be. I think it will be more accurate to be skeptical of these missing values and assume they are as far away as possible.

HAC (single link)

Number of Clusters : 4

Total SSE: 379.412537538

HAC (complete link)

Number of Clusters : 4

Total SSE: 443.312736742

Results

We can see the that there was a significant improvement when using the maximum column values as opposed to assuming unknown values are a distance of 1. My hypothesis was indeed correct.